ARMBench: An Object-centric Benchmark Dataset for Robotic Manipulation

Chaitanya Mitash[1], Fan Wang[1], Shiyang Lu[2], Vikedo Terhuja[1],
Tyler Garaas[1], Felipe Polido[1], Manikantan Nambi[1]

***Abstract*— This paper introduces Amazon Robotic Manipulation Benchmark (ARMBench), a large-scale, object-centric benchmark dataset for robotic manipulation in the context of a warehouse. Automation of operations in modern warehouses requires a robotic manipulator to deal with a wide variety of objects, unstructured storage, and dynamically changing inventory. Such settings pose challenges in perceiving the identity, physical characteristics, and state of objects during manipulation. Existing datasets for robotic manipulation consider a limited set of objects or utilize 3D models to generate synthetic scenes with limitation in capturing the variety of object properties, clutter, and interactions. We present a large-scale dataset collected in an Amazon warehouse using a robotic manipulator performing object singulation from containers with heterogeneous contents. ARMBench contains images, videos, and metadata that corresponds to 235K+ pick-and-place activities on 190K+ unique objects. The data is captured at different stages of manipulation, i.e., pre-pick, during transfer, and after placement. Benchmark tasks are proposed by virtue of high-quality annotations and baseline performance evaluation are presented on three visual perception challenges, namely 1) object segmentation in clutter, 2) object identification, and 3) defect detection. ARMBench can be accessed at http://armbench.com**

I. Introduction

Robotic systems for object handling in warehouses can expedite fulfillment of customer orders by automating tasks such as object picking, sorting, and packing. However, building reliable and scalable robotic systems for object manipulation in warehouses is not trivial. Modern warehouses process millions of unique objects with diverse shapes, materials, and other physical properties. These objects are often stored in unstructured configurations within containers which pose challenges for robotic perception and planning. From 2015 to 2017, the Amazon Robotics Challenge (ARC) helped push the state-of-the-art for robotic systems in a pick-and-place task representative of a warehouse [14], [17]. Nevertheless, the competition could not incorporate challenges of large-scale operations. Fundamental research still needs to be carried out to enable visual perception algorithms such as object segmentation and identification to generalize to a wide variety of unseen objects and configurations. Additional problems (such as defect detection) and metrics (measuring uncertainty in prediction) need to be defined to capture the scale and high-precision requirements of such systems.

[1]Amazon Robotics, MA, USA. {cmitash, fanwanf, terhuja, tggaraas, polidof, mnambi}@amazon.com

[2]Computer Science Department, Rutgers University, NJ, USA. shiyang.lu@rutgers.edu. Work done during a co-op at Amazon Robotics.

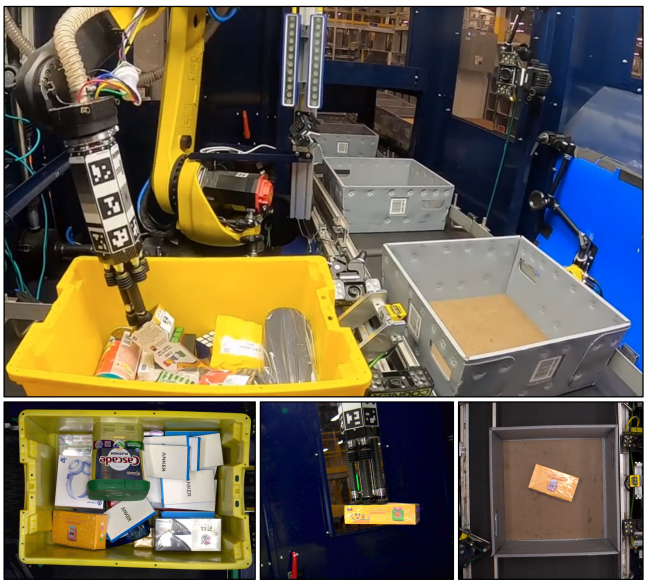

Fig. 1. A large-scale object dataset is collected using a robotic manipulation system operating in an Amazon warehouse. The robotic arm picks one object at a time from the yellow container and places it in a gray tray (top). The dataset contains images for different phases of manipulation i.e., image of objects in the yellow container before picking (bottom-left), during transfer (bottom-mid) and after placement (bottom-right). In addition to sensor data, the dataset also provides high-quality annotations for tasks such as object segmentation, object identification, and defect detection.

Benchmarks and datasets such as ImageNet [56] and MS-COCO [45] have enabled significant performance improvement in computer vision tasks such as image classification and segmentation. No sizeable dataset exists that capture the desired variety of objects, configurations, and interactions in the context of robotic manipulation. Large repositories of 3D shape models [10], [12], [15], [23] enable generating a variety of scenarios with a rich set of annotations in simulation. Nevertheless, they may fail to capture certain physical properties of objects and interactions encountered during manipulation from heterogeneous clutter. Existing real-world datasets [36] operate under closed set assumption with a small number of object types. Such assumptions prevent evaluating algorithms in terms of its generalization capabilities over novel objects which is critical in large scale operations. Additionally, these datasets only deal with static scenes with objects in near perfect conditions and do not consider interactions with a robotic manipulator.

In this paper, we present ARMBench, a large-scale benchmark dataset for a robotic pick-and-place task that captures a wide variety of warehouse objects and configurations. The

dataset comprises images and videos for different stages of robotic manipulation, namely pick, transfer, and place. It includes metadata such as descriptions and reference images for objects in the container. Each pick-and-place activity is also annotated with the identity of the object being manipulated, and the outcome of the manipulation i.e., whether it was successful (a single object was picked and placed) or if it resulted in a defect. The dataset can be used to study different visual perception problems in the context of robotic manipulation. This paper provides novel benchmarks with annotations and baseline performance metrics for:

- **Object Segmentation** including 450,000+ high-quality manual labels for object segments on 50,000+ images. Variations in objects and degree of clutter present a novel challenge for instance segmentation algorithms.
- **Object Identification** presenting an open set object identification and confidence estimation challenge for robotic manipulation. With 190,000+ unique objects in varying configurations, the dataset will be used to benchmark image retrieval and few-shot classification methods with uncertainty estimation.
- **Defect Detection** with manually assigned labels for rare, but costly, robot-induced defects such as multi-object-pick and packaging defects. The dataset contains 19,000+ images and 4,000+ videos of activities with defects, and 100,000+ activities without defects.

II. Related Work

A. Benchmarking in Robotic Manipulation

A recent benchmarking effort for robotic manipulation [9] considers challenges in mechanical design, grasp planning and deformable object manipulation but does not focus on the complexities of underlying perception tasks. An annual competition [36] benchmarks performance of relevant perception algorithms such as object detection, segmentation and 6D pose estimation over a collection of datasets [8], [35], [66]. The Amazon robotics challenge [18], [24] initiated the development of other relevant datasets [55], [41], [68]. While these datasets present interesting challenges in terms of variety of configurations and large occlusions, they are limited in terms of the number of object instances with a maximum of 42 unique objects.

B. Object Segmentation

Object instance segmentation refers to simultaneously predicting pixel-level instance-mask and corresponding class labels. The technique has been widely applied in autonomous driving [69], [21], [54], video surveillance [29], [57], and robotics [67], [19], [43]. The introduction of large-scale labeled dataset such as MS-COCO [45], PASCAL VOC [25], and Cityscapes [16] has significantly advanced the state-of-the-art in detection and segmentation, particularly for common object categories from WordNet [27]. These datasets serve as a standard benchmark for evaluating computer vision models but are not representative of objects a robotic manipulator would interact with. Representative datasets such as the MVTec D2S dataset [28] are limited in size and diversity of objects. To obtain data at scale, a common strategy is to generate synthetic data with physics simulators and rendering tools [20], [48], [32], [67], [63]. Nevertheless, synthetic datasets are limited by the availability of high-quality 3D object models, and inherently carry a *sim2real* gap [63]. This work introduces a real-world, large-scale dataset for object segmentation that captures a wide variety of objects and configurations relevant to robotic manipulation.

C. Object Identification

Object identification refers to the task of exactly identifying the object specified by an image segment. In an open-set setting it is often posed as an image retrieval problem i.e., given a query image and a database of candidate images, rank them according to their similarity to the query image. Various approaches have been used to tackle this problem such as aggregating pre-defined local features [59], [51], [37], [60], computing similarity metrics over features derived from large-scale image classification training [4], [62], [30], and metric learning with pairs of matching and non-matching images [53], [61]. Common benchmarks for image retrieval consider landmark datasets [51], [52], [65] and retail datasets such as DeepFashion [47], [31], Online Products dataset [49], RPC [64], RP2K [50], Products-10K [5], and AliProducts [13]. The product images in these datasets are online store images, customer images or photos from retail stores. Alternatively, the images in the proposed dataset are representative of how objects are stored in a warehouse, with different types of packaging and in cluttered configurations. The robotic manipulation context not only provides a unique set of challenges for object identification in terms of occlusion and viewpoint variations but also imposes stringent requirements in terms of precision.

D. Defect Detection

Few image datasets exist for objects with defects. Prior research on visual defect detection has focused on surface defects for individual objects such as LED chips[44], fabrics [58], and metals [6]. The DAGM 2007 Competition dataset [2] comprises 6,900 synthetic images with six different types of surface defects. The MVTec Anomaly Detection dataset comprises 5,354 color images corresponding to 15 object and texture categories with 70 different types of defects such as scratches, dents, contaminations, and structural changes [7]. This is the first dataset to capture defects in the context of robotic manipulation.

Datasets for defect detection in videos primarily focus on anomaly detection methods for events such as throwing objects, loitering, running [46], crowded scenes [42], and anomalous pedestrian patterns [40]. These datasets contain a limited number of videos (10-50) per activity. Video classification datasets exists in the domain of sports activities [38], human actions [39], and holistic video understanding [22]. Similar to existing research on using videos for understanding context and actions, videos can be used to understand events in robotic manipulation process such as

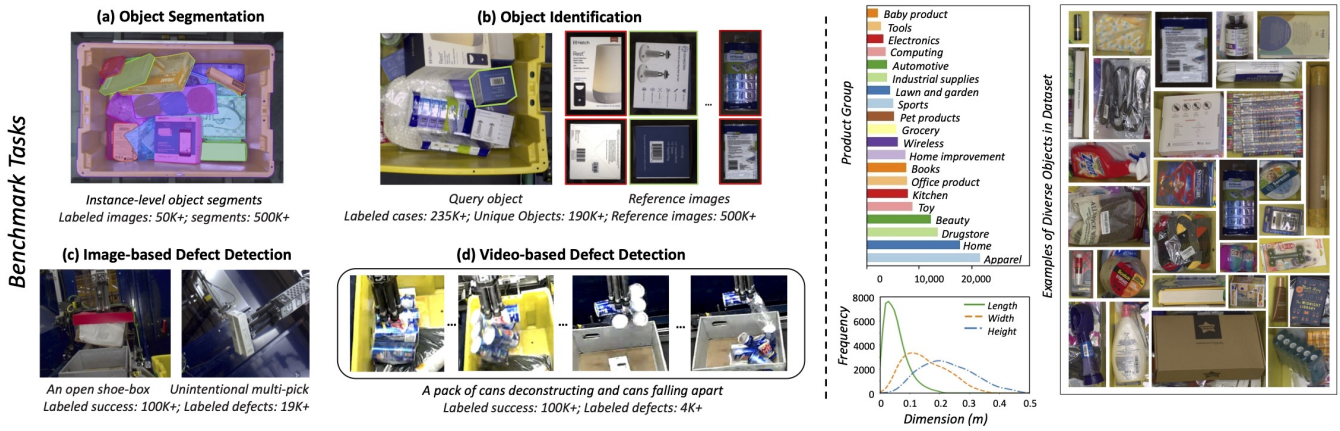

Fig. 2. (left) Benchmark tasks and annotation statistics on the ARMBench dataset. (right) Distribution of product-groups and object dimensions for 190,00+ unique objects in the dataset.

successful and defective activities. There exists no large-scale video datasets for this purpose.

III. ARMBench Dataset

The ARMBench dataset presents: 1) a collection of sensor data acquired by a robotic manipulation workcell performing pick-and-place operation, 2) metadata and reference images for objects in containers, 3) a set of annotations acquired either automatically, by virtue of the system design, or via manual labeling, and 4) tasks and metrics to benchmark perception algorithms for robotic manipulation. Fig. 2 illustrates the benchmark tasks and variety of objects captured in the dataset. The dataset captures diversity in objects with respect to Amazon product categories as well as physical characteristics such as size, shape, material, deformability, appearance, fragility, etc.

The data collection platform is a robotic manipulation workcell performing pick-and-place operation in a warehouse [1]. The workcell contains a robotic arm mounted with a vacuum-based end-effector. It is presented with a heterogeneous collection of objects placed in unstructured configurations within a container (storage tote). The robotic arm is tasked with picking one object at a time (singulation) and place it on moving trays until the container is empty. The empty container ejects the workcell and is replaced by a new container. While the operation is completely autonomous, it includes a human-in-the-loop to monitor the status of each pick-and-place activity, annotate, and resolve any defects during manipulation. Multiple imaging sensors are placed in the workcell to facilitate and validate the pick-and-place operation. Following is a list of sensor data (Fig. 1) associated with each pick activity:

- Pick-image: A 5 MP camera is used to capture a top-down image of the container.
- Transfer-images: Multiple 5 MP cameras are placed on different sides in the workcell to capture the moving object from different viewpoints.
- Place-image: A top-down view of the object is captured once it is placed on the tray.
- Video: A camera is mounted to capture 720p videos of pick-and-place manipulation processes at 30 FPS

Additionally, the following metadata (Fig. 2 (b)) is available by virtue of a warehouse tracking system:

- Container-manifest: A list of objects present in the container along with data such as product description, coarse dimensions, and weight.
- Reference images: One or more images of objects from previous operations within the warehouse.

The sensor data and metadata were consumed by perception algorithms required to autonomously operate the robotic workcell. Benchmarking against these algorithms would not only optimize a manipulation task such as the one used for data collection but also enable more complex and intentional manipulation. This work considers a subset of such perception tasks namely object segmentation, object identification, and defect detection. These are critical not only to make informed grasping and motion decisions but also to track the state of the objects and containers within the warehouse. The following sections will describe these tasks and present the challenges using annotations, baseline algorithms, and evaluation metrics.

IV. Object Segmentation

The object instance segmentation task is to identify and delineate distinct objects stored in containers in a warehouse. In the context of robotic object manipulation, instance segmentation is used to inform downstream robotic processes such as grasp generation, motion planning, and placement. Accuracy of instance segmentation can have an impact on picking success, object identification, and defects introduced in the process. For example, under-segmentation can result in picking multiple objects at a time, while over-segmentation can result in a bad choice of grasp leading to damage or dropping of objects. Fig. 3(a) shows manually annotated object segments on the pick-image. Presence of deformable and transparent objects in clutter makes the task challenging.

Our object instance segmentation dataset contains 50K+ images of objects stored in containers in a warehouse with

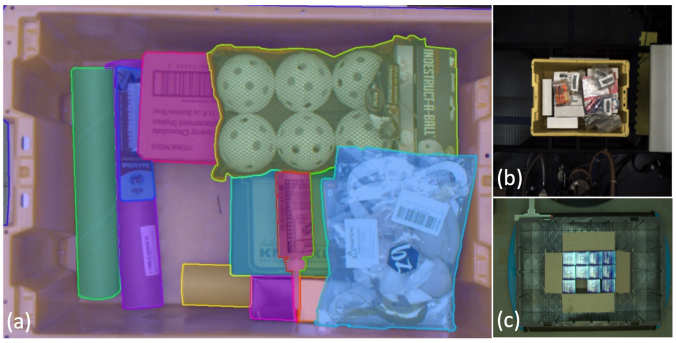

Fig. 3. (a) Segmentation annotation overlaid on an image from *mix-object-tote*. Each identifiable item is segmented regardless of its size and occlusion. Multiple objects in the same package are considered as one object and is delineated by the boundary of the package. In particular, items wrapped in transparent packaging are segmented by the peripheral of the package, although other products may be seen through them. (b-c) Example images from *zoomed-out-tote-transfer-set* and *same-object-transfer-set* subsets representing variations in background, scale, and clutter.

500K+ annotations. The annotations include instance-level segmentation masks and bounding box for two classes (object and container). Technicians with task-specific training generated high-quality annotations for object boundaries and object class which are verified by two additional quality assurance technicians.

We divide the object segmentation dataset into three subsets. The primary set, *mix-object-tote*, comprises 44,253 images and 467,225 annotations of objects in yellow and blue storage totes. The totes contain a heterogeneous clutter of objects with an average of 10.5 object segments (ranging from 1 to 50 segments) in each image. The other two subsets, namely *zoomed-out-tote-transfer-set* and *same-object-transfer-set* (Fig. 3(b) and (c)) enable us to understand the impact of variation in data distribution. The *zoomed-out-tote-transfer-set* subset with 5,837 images and 43,401 annotations captures images of containers from a different warehouse. It poses a transfer learning challenge due to significant differences in background, scale, and object distribution. The *same-object-transfer-set* subset contains 3,323 images and 12,664 annotations. It captures a common and visually challenging scenario in warehouses where multiple instances of the same object are tightly packed in a container.

To establish a performance baseline, we trained Matterport's implementation of Mask R-CNN [3], [33] with ResNet-50 backbone [34] on the *mix-object-tote* dataset. Default training schedule (for MS-COCO) and hyper-parameters were used along with a train-valid-test split of 0.7:0.15:0.15. Table I shows the results for our baseline experiment. Mean average precision (mAP) for a threshold of 0.5 (mAP_{50}) and 0.75 (mAP_{75}) are used to evaluate the performance of the baseline model on test set.

We observe that applying model weights trained on *mix-object-tote* to the *zoomed-out-tote-transfer-set* ($mAP_{50} = 0.25$) and *same-object-transfer-set* subsets ($mAP_{50} = 0.11$) yields poor results. While techniques like transfer learning can improve performance on a new scenario when a reasonable amount of domain-specific labeled data is available, labeling specifically for each variation is time-consuming, if feasible at all. The ultimate goal is to readily transfer segmentation to new scenarios with minimal additional annotations.

TABLE I

MASK R-CNN PERFORMANCE FOR OBJECT SEGMENTATION TASK. THE MODEL WAS TRAINED ON *mix-object-tote* DATASET

	mix-object-tote	zoomed-out-tote-transfer-set	same-object-transfer-set
mAP_{50}	0.72	0.25	0.11
mAP_{75}	0.61	0.19	0.10

We observe that segmentation performance for our baseline model has a strong correlation to the level of clutter. Fig. 4 shows that the performance drops significantly as the number of ground-truth object instances increases in the image. The mAP_{50} score drops sharply from 0.95 when the tote has one to five object instances to a low of 0.38 when there are more than 26 object instances in the image. This motivates developing algorithms that are robust against clutter and occlusion to further improve object segmentation performance.

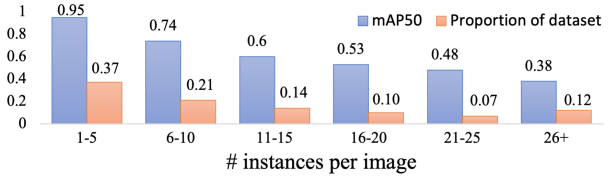

Fig. 4. Performance on *mix-object-tote* with varying degree of clutter.

V. OBJECT IDENTIFICATION

Object identification (ID) is the task of exactly identifying an image segment as one of the objects within a database. In the robotic manipulation context, this task is applicable both before and after picking the object. In the pre-pick stage, identifying an object segment within the tote allows accessing any stored models or attributes of the object from past experience which can be used for manipulation planning purposes. In the post-pick stage, ID has access to the segment of the object being manipulated both within the tote as well as when it is attached to the robotic arm. Accurately identifying which object is being transferred from one container to another is critical to tracking the object within a warehouse, thereby maintaining a container-manifest. This also allows posing ID as an image retrieval challenge. The pick and transfer images with segments of the target object (acquired using an instance segmentation algorithm) are treated as *query images*. The reference images for all objects in the container manifest are treated as *gallery* images. The challenge is to compare the *query images* to *gallery images* to find a match.

The benchmark dataset for this task contains 235K+ labeled pick activities corresponding to 190K+ unique objects. Each pick activity comprises one query image from the pick scene and up to three query images from the transfer phase. A ground-truth ID annotation is automatically acquired using

multiple barcode scanners that are placed in the workcell. In cases where no barcode is scanned, a human operator manually scans the barcode of the object after it ejects the workcell. Pick activities are accompanied by a set of reference images for objects in the container manifest. These images are captures of the object from previous operations within the warehouse. While up to six reference images are sampled per object, reference images are not available for some objects. Such cases are representative of scenarios when a new object is introduced into the warehouse. To tackle such cases, an ID algorithm needs to model some notion of confidence and prevent false-positive prediction. A *test* set is sampled for evaluating baseline algorithms on the dataset. This set contains 50,000 pick activities where at least one reference image is available for the picked object. Another set (*test-uncertainty*) is derived from the *test* set by ignoring reference images of the picked object for 20% of the cases. This set is used to evaluate the behavior of ID on novel objects coming into the warehouse.

TABLE II
EVALUATING TOP-K OBJECT RETRIEVAL

recall@k (pre/post-pick)	k=1	k=2	k=3
ResNet50-RMAC [62]	71.7 / 72.2	81.9 / 82.9	87.2 / 88.2
DINO-ViTS [11]	77.2 / 79.5	87.3 / 89.4	91.6 / 93.5
test-uncertainty			
ResNet50-RMAC [62]	57.7 / 58.0	65.7 / 66.5	70.2 / 70.7
DINO-ViTS [11]	61.9 / 63.6	69.9 / 71.6	73.3 / 74.8

Table II shows results of object retrieval with baseline algorithms on the two sets. For the first baseline, a 512d image descriptor is extracted from a ResNet50 backbone via aggregating features [62] pre-trained for classification on ImageNet dataset [56]. The second baseline utilizes a 384d feature vector pre-trained via self-supervision [11] on a vision transformer. A cosine similarity is computed between feature embeddings for *query* and *gallery* images to get the closest match. Evaluation is performed both over pre-pick (pick image only) and post-pick (pick and transfer images) scenarios. Although transfer images are significantly different in terms of presentation to reference images, they provide multiple views of the object which improves the overall retrieval rate. Fig. 5 shows some of the challenges associated with ID on this dataset. Large variations in appearance for the same object and the similarities between different objects makes the dataset challenging. The challenge increases with the size of container manifest as seen in Fig. 6(left). Fig. 6(right) shows the precision-recall curve obtained based on a rank-ratio confidence metric computed as $(1 - \frac{c_2}{c_1})$, where c_1, c_2 are softmax probabilities corresponding to the first and second ranked objects. The plot highlights recall rates at high precision values as mis-identifications can lead to costly scenarios, such as an object getting lost within the warehouse. While methods like contrastive learning over the training set can improve the top-1 retrieval rate, achieving a high recall rate within the precision constraints would require methods to perform uncertainty estimation and leverage additional modalities such as text and dimensions.

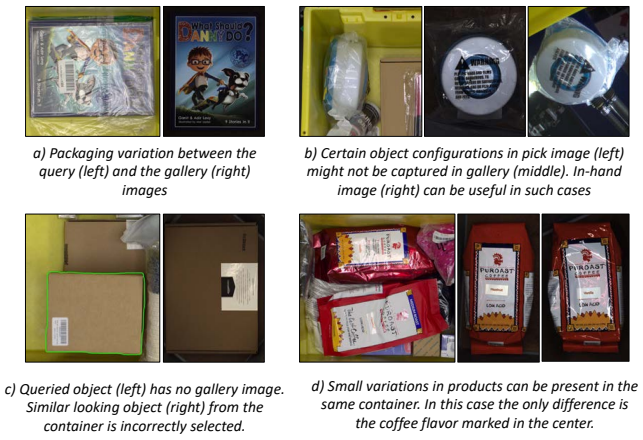

Fig. 5. Challenging cases for Object Identification

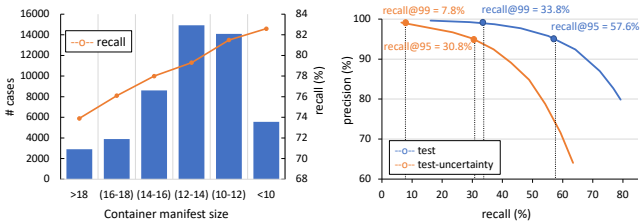

Fig. 6. Container manifest size (left) is indicative of the number of images that the algorithm needs to select from. The precision-recall curve (right) shows the need for a confidence model to prevent false-positive predictions.

VI. DEFECT DETECTION

The defect detection task is to identify if a robotic manipulation activity resulted in a defect. Two types of robot-induced defects are included in the dataset: 1) *multi-pick*, and 2) *package-defect*. *Multi-pick* is used to describe activities where multiple objects were picked and transferred from the source container to the destination container. *Package-defect* is used to describe activities where the object packaging *open*ed and/or the object separated into multiple parts (*deconstruction*). Two subclasses, *open* and *deconstruction*, are defined for package-defect. Fig. 7 shows examples of multi-pick and package-defect in our dataset. Multi-pick are often observed when there is a high degree of clutter, there are multiple instances of the same object, or when objects of significantly different sizes are placed together. Fig. 7 (a-c) shows package-defect on a variety of objects. Defects on deformable objects like plastic bags can be challenging for visual detection.

Our dataset comprises 19,303 images of objects from multiple viewpoints (Transfer-images) and 4,070 videos of pick-and-place activities that resulted in a defect. Videos are excluded from our dataset for multi-pick defect as such defects are not observable along specific viewpoints. Multi-view Transfer-images are best suited to detect multi-pick defect. On the other hand, *open* and *deconstruction* defects can happen at any time during an activity. As a result, they are best captured using videos. The dataset includes 100,000

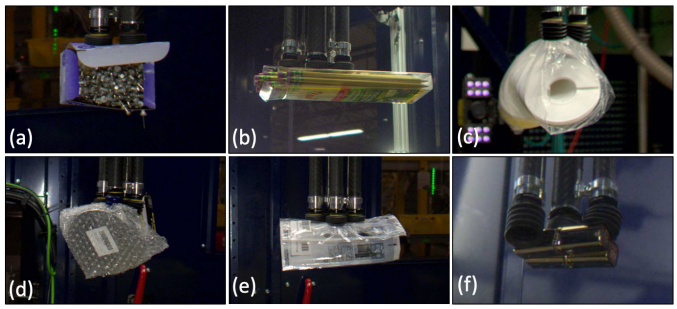

Fig. 7. Multi-view images in the defect detection dataset showing (a)–(c) *package-defect* and (d)–(f) *multi-pick* defect for different types of objects.

images of objects and videos of activities that do not have any defects and are defined as *nominal* activities. Tables III and IV shows the distribution of defect types in our dataset. In addition to Transfer-images, the dataset includes Pick-image and Place-image that provide context for an activity.

A two-step process was used to annotate data. A technician operating our system labeled each activity as successful/nominal (a single object transferred from the source container to the destination container), *multi-pick*, *open*, or *deconstruction* defect. Expert annotators verified the annotations for each activity and augmented the annotations for Transfer-images as multi-pick, or package-defect if a defect was observable, and as nominal if no defect was observable in the image. In addition to the defect type, we also provide segmentation polygons for the objects to enable development of models that can benefit from additional attention cues. For video annotations, expert annotators verified the type of package defect, i.e., *open* and *deconstruction*, observed in the video. In addition to the type of defect observed in each video, the index of the first frame where a defect becomes observable is also provided to enable development of real-time defect detection methods.

To establish a baseline for defect detection, we performed two experiments. In the first experiment, we train an image classifier with ResNet-50 [34] backbone, global average pooling, and focal loss for predicting the type of defect observed in the Transfer-images. In the second experiment, we trained a multi-scale vision transformer model (MViT-B) [26] for action classification on videos. Since a defect can be introduced at any time during the manipulation process, we uniformly sampled 32 frames ($\sim$5 FPS) from each video for training. The classification head outputs a two-channel vector that predicts binary classification on two categories: *open* and *deconstruction*. We used a train-test split of 0.7:0.3 for multi-pick and package-defects. The nominal category in the train set was downsampled to match the size of the defect category to compensate for class imbalance. 10,000 samples from the nominal category were added to our test set.

Table III and IV show performance of baseline models for defect detection on images and videos. We used recall and false positive rate (fpr) as metrics to evaluate performance over defect classes. A missed defect (lower recall) is more expensive than classifying a nominal activity as defective (fpr). Results for image defect detection shows that multi-picks are a harder to detect than package-defects. On the other hand, results for video defect detection show that open defects are harder to detect than deconstruction. There is significant scope for improvement in defect detection methods to be effective in warehouses operations which typically require high recall (>0.95) and low fpr (<0.01).

TABLE III

BASELINE FOR SINGLE-VIEW IMAGE DEFECT DETECTION

model	metric	multi-pick	package-defect	combined
ResNet-50 [34]	count	7,813	11,490	19,303
	recall	0.34	0.73	0.57
	fpr	0.05	0.05	0.05

TABLE IV

BASELINE FOR VIDEO DEFECT DETECTION

model	metric	open	deconstruction	combined
MViT-B [26]	count	2,951	2,165	4,070
	recall	0.69	0.79	0.73
	fpr	0.23	0.03	0.13

VII. DISCUSSION AND FUTURE WORK

In this work we introduced ARMBench, a large-scale, object-centric benchmark dataset for robotic manipulation in warehouses. The object segmentation benchmark presents challenges relating to clutter, deformable and transparent packaging as well as the problem of degrading performance with different backgrounds and storage configurations. The identification benchmark presents an open set recognition challenge on a wide variety of objects. Additionally, images of the same object can vary significantly due to differences in configurations and packaging variations while images of two different objects can appear similar. Missing reference images and high precision requirement makes the benchmark well suited to evaluate uncertainty estimation algorithms. Finally, the defect detection benchmark presents a unique set of challenges such as detection of multi-pick, opening, and deconstruction of packages. Annotations and baselines are provided both for a single-shot as well as video-based detection of such events.

Our intention is for this dataset to grow over time with a goal to increase the number of unique objects, environments, and benchmark tasks. Large-scale sensor data and fine-grained attributes of objects will enable learning generalizable representations that could transfer to other visual perception tasks. We further plan to enrich our dataset with 3D data and annotations, and propose new benchmark tasks.

VIII. ACKNOWLEDGEMENTS

We would like to thank the Sparrow [1] team members for deployment and operation of the robotic workcell, Aalekh (Raj) Ray Chaudhury and the Go-AI team for data annotation support and the Item-matrix team for curating the reference image dataset. We would also like to thank Joey Durham, Andy Marchese, Clay Flannigan, Parris Wellman, Jane Shi and Kapil Katyal for their valuable feedback.

REFERENCES

[1] "Amazon introduces sparrow—a state-of-the-art robot that handles millions of diverse products," https://tinyurl.com/2p8h4w7v.

[2] "DAGM 2007," July 2022, [Online; accessed 12. Jul. 2022]. [Online]. Available: https://conferences.mpi-inf.mpg.de/dagm/2007/prizes.html

[3] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on keras and tensorflow," https://github.com/matterport/Mask_RCNN, 2017.

[4] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European conference on computer vision*. Springer, 2014, pp. 584–599.

[5] Y. Bai, Y. Chen, W. Yu, L. Wang, and W. Zhang, "Products-10k: A large-scale product recognition dataset," *arXiv preprint arXiv:2008.10545*, 2020.

[6] Y. Bao, K. Song, J. Liu, Y. Wang, Y. Yan, H. Yu, and X. Li, "Triplet-graph reasoning network for few-shot metal generic surface defect segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[7] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9584–9592.

[8] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," in *ECCV*, 2014.

[9] B. Calli, A. Dollar, M. A. Roa, S. Srinivasa, and Y. Sun, "Guest editorial: Introduction to the special issue on benchmarking protocols for robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8678–8680, 2021.

[10] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[12] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[13] L. Cheng, X. Zhou, L. Zhao, D. Li, H. Shang, Y. Zheng, P. Pan, and Y. Xu, "Weakly supervised learning with side information for noisy labeled images," in *European Conference on Computer Vision*. Springer, 2020, pp. 306–321.

[14] D. Colling, J. Dziedzitz, K. Furmans, P. Hopfgarten, and K. Markert, "Progress in autonomous picking as demonstrated by the amazon robotic challenge," 2018.

[15] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Y. Vicente, T. Dideriksen, H. Arora, *et al.*, "Abo: Dataset and benchmarks for real-world 3d object understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 126–21 136.

[16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016. [Online]. Available: http://arxiv.org/abs/1604.01685

[17] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Analysis and observations from the first amazon picking challenge," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 1, pp. 172–188, 2016.

[18] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. J. Causo, K. K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Analysis and observations from the first amazon picking challenge," *IEEE Transactions on Automation Science and Engineering*, vol. 15, pp. 172–188, 2018.

[19] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, "Segmenting unknown 3d objects from real depth images using mask R-CNN trained on synthetic point clouds," *CoRR*, vol. abs/1809.05825, 2018. [Online]. Available: http://arxiv.org/abs/1809.05825

[20] ——, "Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2019.

[21] B. De Brabandere, D. Neven, and L. Van Gool, "Semantic instance segmentation for autonomous driving," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 478–480.

[22] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhagen, and L. V. Gool, "Large scale holistic video understanding," in *European Conference on Computer Vision*. Springer, 2020, pp. 593–610.

[23] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," *arXiv preprint arXiv:2204.11918*, 2022.

[24] C. Eppner, S. Höfer, R. Jonschkowski, R. Martín-Martín, A. Sieverling, V. Wall, and O. Brock, "Lessons from the amazon picking challenge: Four aspects of building robotic systems," in *IJCAI*, 2017.

[25] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–308, September 2009, printed version publication date: June 2010.

[26] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6804–6815, 2021.

[27] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*, ser. Language, Speech, and Communication. Cambridge, MA: MIT Press, 1998.

[28] P. Follmann, T. Böttger, P. Härtinger, R. König, and M. Ulrich, "MVTec D2S: densely segmented supermarket dataset," *CoRR*, vol. abs/1804.08292, 2018. [Online]. Available: http://arxiv.org/abs/1804.08292

[29] V. Gajjar, A. Gurnani, and Y. Khandhediya, "Human detection and tracking for video surveillance a cognitive science approach," 2017.

[30] N. Garcia and G. Vogiatzis, "Learning non-metric visual similarity for image retrieval," *Image and Vision Computing*, vol. 82, pp. 18–25, 2019.

[31] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5337–5345.

[32] M. Grard, L. Chen, and E. Dellandréa, "Bicameral structuring and synthetic imagery for jointly predicting instance boundaries and nearby occlusions from a single image," *CoRR*, vol. abs/1906.07480, 2019. [Online]. Available: http://arxiv.org/abs/1906.07480

[33] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[35] T. Hodan, P. Haluza, S. Obdrzálek, J. Matas, M. I. A. Lourakis, and X. Zabulis, "T-less: An rgb-d dataset for 6d pose estimation of texture-less objects," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 880–888, 2017.

[36] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, *et al.*, "Bop: Benchmark for 6d object pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.

[37] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3304–3311.

[38] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[39] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[40] F. Landi, C. G. Snoek, and R. Cucchiara, "Anomaly locality in video surveillance," *arXiv preprint arXiv:1901.10364*, 2019.

[41] J. Leitner, A. W. Tow, N. Sünderhauf, J. E. Dean, J. W. Durham, M. Cooper, M. Eich, C. F. Lehnert, R. Mangels, C. McCool, P. T.

Kujala, L. Nicholson, T. T. Pham, J. Sergeant, L. Wu, F. Zhang, B. Upcroft, and P. Corke, "The acrv picking benchmark: A robotic shelf picking benchmark to foster reproducible research," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4705–4712, 2017.

[42] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 1, pp. 18–32, 2013.

[43] H. Liao, T. Inomata, I. Sakuma, and T. Dohi, "3-d augmented reality for mri-guided surgery using integral videography autostereoscopic image overlay," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 6, pp. 1476–1486, 2010.

[44] H. Lin, B. Li, X. Wang, Y. Shu, and S. Niu, "Automated defect inspection of LED chip using deep convolutional neural network," *Journal of Intelligent Manufacturing*, vol. 30, no. 6, pp. 2525–2534, 2019.

[45] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[46] W. Liu, D. L. W. Luo, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[47] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.

[48] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *CoRR*, vol. abs/1703.09312, 2017. [Online]. Available: http://arxiv.org/abs/1703.09312

[49] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4004–4012.

[50] J. Peng, C. Xiao, and Y. Li, "Rp2k: A large-scale retail product dataset for fine-grained image classification," *arXiv preprint arXiv:2006.12634*, 2020.

[51] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.

[52] ——, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[53] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[54] M. Ren and R. S. Zemel, "End-to-end instance segmentation and counting with recurrent attention," *CoRR*, vol. abs/1605.09410, 2016. [Online]. Available: http://arxiv.org/abs/1605.09410

[55] C. Rennie, R. Shome, K. E. Bekris, and A. F. de Souza, "A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place," *IEEE Robotics and Automation Letters*, vol. 1, pp. 1179–1185, 2016.

[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[57] N. O. Salscheider, "Object tracking by detection with visual and motion cues," *CoRR*, vol. abs/2101.07549, 2021. [Online]. Available: https://arxiv.org/abs/2101.07549

[58] J. Silvestre-Blanes, T. Albero-Albero, I. Miralles, R. Pérez-Llorens, and J. Moreno, "A public fabric database for defect detection methods and results," *Autex Research Journal*, vol. 19, no. 4, pp. 363–374, 2019. [Online]. Available: https://doi.org/10.2478/aut-2019-0035

[59] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.

[60] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: aggregation across single and multiple images," *International Journal of Computer Vision*, vol. 116, no. 3, pp. 247–261, 2016.

[61] G. Tolias, T. Jenicek, and O. Chum, "Learning and aggregating deep local descriptors for instance-level recognition," *arXiv preprint arXiv:2007.13172*, 2020.

[62] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.

[63] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *CoRR*, vol. abs/1809.10790, 2018. [Online]. Available: http://arxiv.org/abs/1809.10790

[64] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "Rpc: A large-scale retail product checkout dataset," *arXiv preprint arXiv:1901.07249*, 2019.

[65] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2575–2584.

[66] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *ArXiv*, vol. abs/1711.00199, 2018.

[67] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen object instance segmentation for robotic environments," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1343–1359, 2021.

[68] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1386–1383.

[69] Z. Zhang, S. Fidler, and R. Urtasun, "Instance-level segmentation for autonomous driving with deep densely connected mrfs," 2016.